面向小样本关系抽取的自适应胶囊网络 *

张晓明 [1a]，窦全胜 [1b,2†]，陈淑振 [1b]，唐焕玲 [1b, 2]

(1. 山东工商学院 a. 信息与电子工程学院; b. 计算机科学与技术学院，山东 烟台 264000; 2. 山东省高等学校未来智能计算协同创新中心，山东 烟台 264000)

摘　要：小样本关系抽取任务是自然语言处理中的热点问题，旨在使用低成本的标注数据训练关系抽取模型。目前广泛使用的原型网络存在类原型表达不准确、不完整等问题。为了克服该问题，提出一种自适应胶囊网络(Adaptive Capsule Network，ACNet)，ACNet 借助胶囊网络的归纳能力生成类原型，并在此基础上对动态路由过程进行评估，使其面对不同样本能自适应调整网络参数。同时，在 ACNet 中引入一种记忆迭代机制，帮助模型快速确定类表示。在小样本关系数据集 FewRel 上进行实验验证得出，ACNet 能够有效处理小样本关系抽取任务。

关键词：关系抽取；小样本；自适应；胶囊网络；动态路由

中图分类号：TP391　　doi: 10.19734/j.issn.1001-3695.2021.12.0702

Adaptive capsule network for few-shot relation extraction

Zhang Xiaoming[1a], Dou Quansheng[1b,2†], Chen Shuzhen[1b], Tang Huanling[1b,2]

(1.*a. School of Information & Electronic Engineering*, *b. School of Computer Science & Technology*, *Shandong Technology & Business University*, *Yantai Shandong* 264000, *China*;2. *Shandong Future Intelligent Computing Collaborative Innovation Center*, *Yantai Shandong* 264000, *China*)

Abstract: The few-shot relationship extraction task is a hot issue in natural language processing. It aims to train the relationship extraction model using low-cost label data. The widely used prototype network has some problems, such as inaccurate and incomplete expression of class prototypes. This paper proposed an Adaptive Capsule Network (ACNet) to solve this problem. ACNet generates a class prototype with the inductive capability of the capsule network. On this basis, the dynamic routing process is evaluated so that it can adjust network parameters adaptively to different samples. At the same time, a memory iteration mechanism is introduced in ACNet to help the model determine the class representation quickly. Experiments on a few-shot relational dataset FewRel show that ACNet can handle few-shot relational extraction tasks.

Key words: relationship extraction; few-shot; adaption; capsule network; dynamic routing

0　引言

由自然语言构成的文本数据是当前大数据的重要组成部分，在文本中，如人名、地名等具有特殊意义的词汇被称为实体(Entity)，实体之间通常存在着某种关系，如语法关系、语义关系等．所谓实体关系抽取(Entity Relation Extraction)，就是指在实体识别的基础上，对实体之间存在的上述关系进行有效的识别和判断。实体关系抽取是正确理解文本语义的关键，也是文本挖掘和信息抽取的关键基础性任务，其效果对文本摘要、自动问答[1]、机器翻译[2]、语义网标注、知识图谱[3]等自然语言处理下游任务有着重要的影响，一直是自然语言处理(Natural Language Processing, NLP)研究领域的重要研究内容和热点研究问题。

近年来，深度学习的兴起为关系抽取任务提供了新的解决方案，这方法多采用监督学习方式，其效果对样本数据存在较强依赖。然而在现实场景中，数据量往往难以满足大规模深度网络训练的需要。为了避免数据收集带来的人力和时间成本，一些学者提出小样本学习(Few-Shot Learning，FSL)[4]的概念，探索深度学习模型在小样本条件下泛化能力。关于小样本学习的早期研究，多集中于数据增强[5]和正则化技术，通过这两种技术来缓解由数据稀疏引起的过拟合问题。有学者受人类学习过程的启发，提出了元学习[6](Meat-Learning，ML)的概念。ML 通常将小样本学习的训练过程分解为若干个元任务，元任务在不同 mini-batch 之间切换并从中提取一些可迁移的知识。因此，Few-shot 模型只需使用少量标记样本可以对新类别进行分类。

然而，现有的小样本学习方法仍面临许多重要问题，包括强先验方法的弱移植性[7]、复杂的任务梯度转移[8]、以及微调目标问题[9]。Snell[10]和 Sung[11]等人提出的方法结合了非参数方法和度量学习，为其中一些问题提供了解决方案。非参数方法的优势在于能够快速吸收新样本，且模型只需学习样本的表示和度量，这在一定程度上避免了过拟合。但同一类中的实例是相互关联的，并且有它们的统一分数和特定分数。在之前的研究中，类级表示多通过简单地求和或平均支持集样本特征来计算。鉴于实例样本的多样性，这种方法所获得的类级表示往往会受到不同形式样本的噪声影响。且现有小样本学习算法大多不会对支持集进行微调。当增加支持集的大小时，数据扩充带来的改进也会被更多的样本级噪声削弱。

2017 年 Sabour[12]等人提出了胶囊网络，该网络具有解决类表达问题的潜力，胶囊网络将样本向量封装为“胶囊”，并通过非参数的动态路由算法(Dynamic Routing，DR)对部分和整体之间的内在空间关系进行编码。类似的，在小样本任务中，将样本视为部分，类视为整体，DR 算法编码的类表示更具代表性。

收稿日期：2021-12-08；**修回日期：**2022-03-07　**基金项目：**国家自然科学基金资助项目(61976125，61976124)

作者简介：张晓明(1997-)，男，山东临沂人，硕士研究生，主要研究方向为深度学习、人工智能、自然语言处理；窦全胜(1971-)，男(通信作者)，山东烟台人，教授，硕导，博士，主要研究方向为机器学习、人工智能、演化计算(li_dou@163.com)；陈淑振(1997-)，男，山东济宁人，硕士研究生，主要研究方向为深度学习、自然语言处理；唐焕玲(1970-)，女，山东烟台人，硕导，博士，主要研究方向为机器学习、人工智能、数据挖掘.

本文在胶囊网络的基础上，提出一种自适应胶囊网络(Adaptive Capsule Network，ACNet)，旨在从少量支持集样本中发掘样本类别的广义类表示。在胶囊网络中，DR 算法的路由次数决定了部分到整体的层次关系，现有的路由算法对所有样本使用相同的路由次数，面对复杂的实例环境难以有效做到类归纳，ACNet 对胶囊网络中的动态路由算法进行改进，提出一种自适应归纳算法(Self-adaption Inductive Algorithm，SIA)，SIA 通过评估路由算法在实例样本上的执行过程，为不同类样本分配相应的路由次数，实现路由参数的自适应调整。同时，为了降低不同样本所带来的噪声干扰，ACNet 引入了一种可训练的记忆模块帮助路由过程快速确定类表示，记忆模块中包含不同类的类特征，这些类级表示作为模型的学习经验，有效缓解了样本量过少带来的路由过程不准确问题。

综上所述，本文主要贡献包括：

a) 提出一种自适应胶囊网络。该网络将记忆保存机制与动态路由算法结合，能够快速适应支持集样本，并在小样本场景中有效归纳样本类表示。

b) 提出一种自适应归纳算法。该算法在动态路由的基础上引入一种路由过程的评估机制，使模型能够针对不同样本自适应的分配路由参数，缓解因样本多样性导致的类特征难聚合、表达不完善等问题。

将本文方法在 FewRel 数据集上进行实验，实验结果证明了本文研究方法的有效性，对小样本场景下的关系抽取任务，具有较强的指导意义和应用价值。

1　问题描述

为了论述上的便利和准确，以下就实体关系抽取任务给出形式化描述。

设 $W=\{w_i\}_{i=1}^{m}$ 和 $E=\{e_i\}_{i=1}^{n}$ 分别为单词符号集合和实体标记集合，W 上的文本 t 可视作 W 中元素构成的有限长度序列：$t=w_1^t w_2^t \ldots w_h^t$，在文本中，具有特殊语义的一个或一组单词符号被称为实体，每个实体于 E 中标记对应：$t=w_1^t\cdots\underbrace{w_i^t\cdots w_j^t}_{e_p}\cdots\underbrace{w_u^t\cdots w_v^t}_{e_q}\cdots w_h^t$ 其中，$w_i^t\cdots w_j^t$ 和 $w_u^t\cdots w_v^t$ 分别对应于实体 e_p 和 e_q。实体识别和标注是 NLP 处理的另一项重要任务，在此不做讨论，本文在实体识别的基础上进一步分析语句中两个实体之间的关系。

实体间的联系构成了一个关系集合，记作 $R=\{r_j\}_{j=1}^{k}$，文本 t 中的一对实体 e_p 和 e_q 的关系可用三元组 $\langle e_p,r,e_q\rangle$ 表示，其中 $r\in R$ 为目标关系集的一个元素。实体关系抽取任务的目的是从自然语言文本中抽取出这样的关系三元组，为更深入的文本挖掘和理解奠定基础。以句子“London is the capital of the UK”为例，其中“London”和“UK”为两个反映地名的实体，两个实体间存在语义关系：“capital of”， <London ,Capital of，UK>即为一个实体关系三元组。

小样本实体关系抽取任务是针对某些任务领域关系样本数量稀少，无法开展大规模模型训练的情况。在该场景下，给定关系集合 R 和只包含少量样本的支持集 S，要求模型能够准确预测查询样本语句 x 中实体对 e_p 和 e_q 间的关系。其中，支持集 S 和查询集 Q 均通过对数据集 D 采样获得，即在数据集 D 中随机选择 C 个类别，并在每个类别中随机选择 K 个样本构成支持集：$S=\{s_{c,k}\}$，$c=1,\ldots,C, k=1,\ldots,K$。另外在 C 个类别的其余样本中随机选择 R 个样本构造查询集：$Q=\{s_q\}, q=1,\ldots,R$。这种构建支持集与查询集的任务方式也被称作 C-way K-shot。在小样本学习中，支持集中的实例数量(即 C*K 个)通常很少，关系分类模型需要在支持集的少数示例中学习样本特征，并预测查询实例 x 的关系类别。

2　自适应胶囊网络模型(ACNet)

本文提出了用于小样本关系分类的自适应胶囊网络模型(ACNet)，模型共包含基类数据 s_i^{base}、支持集 $s_{c,k}$ 和查询集 s_q 三个输入，其中 s_i^{base} 由训练集 D_{train} 生成，$s_{c,k}$ 和 s_q 以 c-way-k-shot 任务为标准在 D_{train} 中随机抽取获得。ACNet 模型四个模块，具体如图 1 所示。

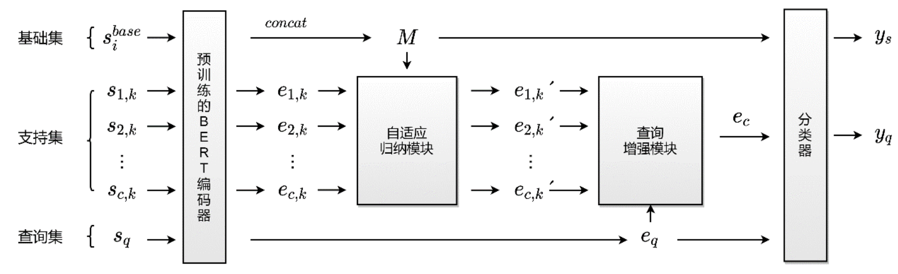


图 1　自适应胶囊网络(ACNet)结构示意图

Fig. 1　Structure diagram of Adaptive Capsule Network (acnet)

1) 编码模块：采用预训练的 BERT[13]模型分别对基类数据 s_i^{base}、支持集 $s_{c,k}$ 和查询集 s_q 进行编码，分别得到基类向量 e_i、支持向量 $e_{c,k}$ 和查询向量 e_q。

2) 自适应归纳模块(Adaptive Induction Module，AIM)：采用自适应归纳算法(SIA)对支持集向量 $s_{c,k}$ 进行归纳，SIA 在动态路由算法的基础上，加强对路由过程的评估，针对不同的支持向量自适应调整路由参数，经过多次路由迭代，获得支持集的嵌入向量 $e_{c,k}'$。

3) 查询增强模块(Query Enhancement Module，QEM)：复用自适应归纳算法，在 $e_{c,k}'$ 的基础上对查询向量 e_q 进行调整，获得包含查询信息的类表示 e_c，用于后续 e_q 的分类。

4) 分类模块：采用余弦相似度量计算查询向量 e_q 与类表示 e_c 的匹配分数，预测查询向量类别。

2.1　编码模块

选择预训练的 BERT 模型作为编码工具，其模型架构是基于 Transformer[14]的多层双向编码器。在关系语句中插入[cls]和[sep]作为开头和结尾的标识符号，并使用[cls]输出的 d 维向量作为给定实例关系语句 w 的向量表示，整个过程可以表示为 $e=E(w|\theta)$，其中 θ 为 BERT 的模型参数。预训练的 BERT 模型提供了强大的上下文相关句子表示，可用于各种目标任务，并且适用于小样本关系分类。

使用 wiki 文本对 BERT 模型进行预训练，为了能够适应小样本关系分类任务，在训练集 D_{train} 中随机取 C_{base} 个类别的数据组成基类样本 $\{s_i^{base}\}_{i=1}^{C_{base}}$ 用于对 BERT 编码器进行微调。对于每个输入 s_i^{base}，编码器 $E(s_i^{base}|\theta)$ 输出 d 维向量 e_i。同时有矩阵 $M=[e_1,e_2\cdots e_i\cdots e_{C_{base}}]$，M 作为记忆矩阵为每个基类样本保存一个记忆特征向量，为了保证记忆特征的有效性，$E(s_i|\theta)$ 和 M 都将在模型训练过程中进一步调整，具体细节将在 2.2 节中阐述。

2.2　自适应归纳模块(Adaptive Induction Module，AIM)

AIM 旨在利用记忆矩阵 M 对支持集进行调整，将多个记忆特征和支持向量输入到 AIM 中，经过自适应归纳算法

(SIM)获得每个支持向量的嵌入向量，DR 算法可以实现多向量到单一向量归纳的功能，SIM 在此基础上改进了路由过程的评估方法，并依照其评估结果自适应的为不同样本分配路由参数，获得的嵌入向量能够有效整合记忆特征中的信息，适应支持集的能力更强。

具体地，支持集中的实例首先被 BERT 编码为样本向量 $\{e_{c,k}\}_{k=1}^{K}$，然后输入到 AIM 中处理。在给定记忆矩阵 M 和支持样本向量 $e_{c,k}$ 的情况下，AIM 旨在使用记忆矩阵 M 来调整支持向量，整个过程可以概括为：$e_{c,k}'=AIM(M,e_{c,k})$。

在胶囊网络中，存在 l 和 l+1 两个胶囊层，低级胶囊分布在胶囊层 l 上，通过动态路由算法将多个低级胶囊以加权和的方式路由到胶囊层 l+1 上，并获得高级胶囊 v_j，鉴于胶囊网络部分到整体的编码特性，在小样本学习中，将低级胶囊视为样本，而高级胶囊代表样本类别特征，因此在本文中，对于输入 AIM 的每个 $m_i \in M$ 和 $e_{c,k}$ 进行标准矩阵转换，并应用 squash 函数[12]进行归一化：

$$\hat{m}_i = squash(W_j m_i + b_j) \tag{1}$$

$$\hat{e}_{c,k} = squash(W_j e_{c,k} + b_j) \tag{2}$$

这里的转换权重 W_j 和参数 b_j 在输入中共享，需要在网络中学习得到，*squash* 函数为非线性压缩函数，目的是在保持向量方向不变的条件下，将其长度压缩至区间[0,1]内，函数通式如下：

$$squash(x)=\frac{\|x\|^2}{1+\|x\|^2}\frac{x}{\|x\|} \tag{3}$$

在胶囊网络中，高级胶囊 v_j 由低级胶囊的加权和计算得到：

$$v_j=\sum_{i=1}^{n}(c_{ij}+p_{ij})\hat{m}_i \tag{4}$$

c_{ij} 为不同等级胶囊间的路由权重，$p_{ij}=tanh(PCCs(\hat{m}_i,\hat{e}_{c,k}))$，PCCs[15]用于度量基类类特征 $\hat{m}_i$ 和支持向量 $\hat{e}_{c,k}$ 间的相似程度，具体如式(5)所示。

$$PCCs=\frac{Cov(x_1,x_2)}{\sigma_{x_1}\sigma_{x_2}} \tag{5}$$

其中，Cov 为协方差，σ_{x_1} 和 σ_{x_2} 分别为向量 x_1 和 x_2 的标准差，由于 PCCs 的值分布在[-1,1]区间内，因此可用于增强或惩罚路由参数 c_{ij}。

在传统 DR 算法中，路由次数 r 的大小决定了不同胶囊层之间的层次关系，在强监督环境中，网络在开始训练时往往需要选取合适的 r 值，并将其应用在所有样本上，但在小样本 RE 任务中，数据规模不足且关系实例多样复杂，不同样本达到收敛要求的路由次数也不相同，固定的 r 值难以满足任务需要。因此，为了评估路由过程在小样本实例上的性能，并实现路由次数的自适应调整，本文提出一种**自适应归纳算法**(Self-adaption Inductive Algorithm，SIA)来解决此问题。SIA 将路由过程描述成最小化负一致性分数(Negative Agreement Score, NAS)的优化问题：

$$\min_{c,v} f(u)=-\sum_{i,j}c_{ij}\langle v_j,\hat{m}_i\rangle \quad \text{s.t.}\quad \forall i,j: c_{ij}>0,\ \sum_j c_{ij}=1. \tag{6}$$

NAS 的目的是将更高的 c_{ij} 值分配给更加接近的 $\langle v_j,\hat{m}_i\rangle$ 对，但鉴于向量的高维特征，NAS 函数的计算一直是一个开放问题。核密度估计(Kernel Density Estimation，KDE)是一种非参数密度估计方法，不需要假设一致对是从参数分布中提取的，这为解决 NAS 问题提供了可能。通过 KDE 将式(6)转换成如下形式：

$$\min_{c,m} f(u)=-\sum_{i,j}c_{ij}k(d(v_j,\hat{m}_i)) \tag{7}$$

其中：*d* 为欧式距离，*k* 为 Epanechnikov[16]核函数：

$$k(x)=\begin{cases}1-x & x\in[0,1)\\ 0 & x\geq 1\end{cases} \tag{8}$$

均值漂移(Mean Shift)[17]是一种基于密度梯度上升的非参数方法，可用于最小化 NAS 函数 $f(u)$ 以解决 KDE 问题：

$$\nabla f(u)=\sum_{i,j}c_{ij}k'(d(v_j,\hat{m}_i))\frac{\partial d(v_j,\hat{m}_i)}{\partial v} \tag{9}$$

确定 c_{ij}^{τ} 的值后，对高级胶囊 $v_j^{\tau+1}$ 进行更新：

$$v_j^{\tau+1}=\frac{\sum_{i,j}c_{ij}^{\tau}k'(d(v_j,\hat{m}_i))\hat{m}_i}{\sum_{i,j}k'(d(v_j,\hat{m}_i))} \tag{10}$$

高级胶囊 $v_j^{\tau+1}$ 中聚合了记忆特征的信息，通过取平均的方式对支持向量 $\hat{e}_{c,k}$ 进行调整：

$$\hat{e}_{c,k}\leftarrow\frac{\hat{e}_{c,k}+v_j}{2} \tag{11}$$

$c_{ij}^{\tau+1}$ 的值可以使用标准梯度下降进行更新：

$$c_{ij}^{\tau+1}=c_{ij}^{\tau}+\alpha\cdot(k(d(v_j^{\tau},m_i))+p_{ij}) \tag{12}$$

其中：α 为步长。$k(d(v_j^{\tau},m_i))$ 为不同等级胶囊间的相似度估计，p_{ij} 为 $k(d(v_j^{\tau},m_i))$ 的修正量，代表记忆特征向量 m_i 与支持向量 $e_{c,k}$ 的相似程度，相似程度越高对应的记忆特征在 c_{ij} 中的比重越大，反之则会削弱不相关的记忆特征占比。通过估计值加修正量的方式更新胶囊权重 c_{ij}，保证生成高级胶囊 v_j 时能够有效聚合记忆特征，并在此基础上实现对支持向量的调整。

为了解决路由算法在实例级样本上不会收敛的问题，SIA 可以根据单个实例样本 NAS 取值自行调整路由迭代次数，详细过程如算法 1 所示：步骤 a) 代表遍历所有记忆特征 m_i 和支持向量 $e_{c,k}$，并在步骤 b) 和 c) 分别对其进行归一化计算，步骤 e) 代表采用加权和的方式计算类特征 v_j 的初值，算法在步骤 f) 进入路由迭代循环，步骤 h) 代表 v_j 的更新过程，具体见式(6)~(10)，步骤 i) 和 j) 表示在获得最新的 v_j 后，路由权重 c_{ij} 和支持向量 $e_{c,k}$ 的调整更新方式，步骤 k) 代表 NAS 函数的取值计算，步骤 l) 为判断语句，比较 NAS 更新前后的差值与阈值 β 大小，若 $|NAS-Last_NAS|>\delta$，则表明 NAS 函数不满足收敛标准，算法进入步骤 o) 和步骤 p)，分别对修正量 p_{ij} 和 NAS 进行更新，并重新返回步骤 f) 进行下一轮的路由迭代，直至 $|NAS-Last_NAS|<\delta$，即 NAS 已具备收敛条件，退出步骤 f) 的 while 循环，进入步骤 q) 输出自适应嵌入向量 $e_{c,k}'$，算法结束。

算法 1　自适应归纳算法(SIA)

输入：超参数 α，β；路由权重 $c_{ij}=1/n$；支持向量 $e_{c,k}$；记忆矩阵 $M=[m_1,m_2,\ldots,m_n]$。

输出：自适应嵌入向量 $e_{c,k}'$。

a): **for all** $m_i, e_{c,k}$ **do**

b): $\hat{m}_i = squash(W_j m_i + b_j)$

c): $\hat{e}_{c,k} = squash(W_j e_{c,k} + b_j)$

d): $p_{ij} = tanh(PCCs(\hat{m}_i,\hat{e}_{c,k}))$

e): $v_j=\sum_{i=1}^{n}(c_{ij}+p_{ij})\hat{m}_{ij}$

f): **While true do**

g): $c_{ij}\leftarrow \mathrm{softmax}(c_{ij})$

h): $v_j\leftarrow\frac{\sum c_{ij}k'(d(v_j,\hat{m}_i))\cdot\hat{m}_i}{\sum_{i=1}^{n}k'(d(v_j,\hat{m}_i))}$

i): **For all** i: $c_{ij}\leftarrow c_{ij}+\alpha\cdot(k(d(v_j,\hat{m}_i))+p_{ij})$

j): **For all** k: $\hat{e}_{c,k}\leftarrow\frac{\hat{e}_{c,k}+v_j}{2}$

k): $NAS=\log(\sum_{i,j}c_{ij}k(d(v_j,\hat{m}_i)))$

l): **If** $|NAS-Last_NAS|<\delta$ **then**

m): **break**

n): else

o): For all $i,k: p_{ij} = tach(PCCs(\hat{m}_i, \hat{e}_{c,k}))$

p): $Last_NAS \leftarrow NAS$

q): Return $e'_{c,k} = v_j$

2.3 查询增强模块(QEM)

为了避免实例多样性所带来的噪声干扰，在上述两个模块获得的查询向量 e_q 以及嵌入向量 $\{e'_{c,k}\}_{k=1}^{K}$ 的基础上，构建查询增强模块。QEM 目的是在嵌入向量中发掘与查询向量的相似部分，以此构造包含查询信息的类级向量。由于 SIA 具有自适应的能力，可以增强相似的嵌入和查询向量，并对不相关的向量权重进行削弱。因此，通过复用 SIA 在支持向量的基础上对查询集向量进行适应调整，并从与查询集更加相关的嵌入向量中得到类级别的向量表示：

$$e_c = \mathrm{AIM}(\{e'_{c,k}\}_{k=1}^{K}, e_q) \tag{13}$$

2.4 相似度分类器

在最后的分类阶段，对基类向量 e_i 和查询向量 e_q 进行分类，获得所有类别的概率分布。传统神经网络分类器是在提取特征向量 $e \in R^d$ 之后，使用内积 $s_k = e^T w_k^*$ 计算每个类别 $k \in [1, K^*]$ 的初始得分，其中 w_k^* 为权重向量，然后使用 softmax 函数计算特征向量在所有 K^* 类上的分类概率。然而，这种方法不再适用于样本中包含新类的小样本学习。本文使用余弦相似度计算原始分类分数：

$$s_k = \tau \cdot \cos(e_i, w_k^*) = \tau \cdot \overline{e}_i^T \overline{w}_k^* \tag{14}$$

其中，$\overline{e}_i$ 和 $\overline{w}_k^*$ 是 l_2 −正则化向量，τ 是可学习的参数。基类向量 e_i 在 C_{base} 个类别上的分类概率为

$$\hat{y}_s = softmax(s_k) \tag{15}$$

在小样本关系分类场景中，将查询向量 e_q 和类表示 e_c 统一输入分类器，得到小样本学习部分每个新类的分类得分：

$$s_{q,c} = \tau \cdot cos(e_q, e_c) = \tau \cdot \overline{e}_q^T \overline{e}_c \tag{16}$$

$$s_q = \{s_{q,c}\}_{c=1}^{C} \tag{17}$$

查询向量 e_q 在 C 个新类的分类概率为

$$\hat{y}_q = softmax(s_q) \tag{18}$$

3 模型学习过程

在小样本关系分类任务中，训练集 D_{train} 和测试集 D_{test} 具有不同的标签空间，即 $R_{train} \cap R_{test} = \varnothing$，每个数据集的样本可表示为 (x, p, y) 的形式，其中表包含 t 个单词的关系实例语句，$p = (p_1, p_2)$ 表示中两个标注实体的位置，为实例语句中实体对间的关系类别标签。

ACNet 模型在训练时采用 Vinyals[18]提出的基于元任务的训练策略。在该策略中，小样本的学习过程被分成了元训练和元测试两个阶段。在元训练阶段，ACNet 模型将面对许多独立的监督任务 T(即元任务)，不同元任务间的类别不完全相同。每个 T 都以 C-way-K-shot 任务为标准，在训练集 D_{train} 上随机构造，包括支持集 S_T 和查询集 Q_T。将 S_T 中的样本输入 ACNet 模型进行训练，并使用 Q_T 中的样本对模型进行测试，面对每个元任务 T，采用如下交叉熵损失推动模型进行训练：

$$L(S_T, Q_T) = -\frac{1}{C}\sum_{c=1}^{C}\frac{1}{R}\sum_{q}^{R} y_q \log(\hat{y}_q) \tag{19}$$

其中，C 代表 S_T 中的样本类别，R 为 Q_T 中的样本数量，y_q 和 $\hat{y}_q$ 分别表示样本真实标签和模型预测标签。在元测试阶段，测试集 D_{test} 的设置与训练时相同。测试集与训练集合具有不同的标签空间，因此，新任务中的样本类别是之前学习过的任务中没有出现的，ACNet 在训练过程中不断学习新的任务，在经过大量的不同任务训练之后，能够更好地处理任务之间的不同并忽略特定任务的特征，在面对新的小样本任务时，具有更强的泛化能力。

4 实验

4.1 数据集、评估指标

本文在小样本关系分类数据集 FewRel[19],[20]上对模型进行评估。FewRel 数据集分为 1.0 与 2.0 两个版本，FewRel1.0 使用 Wikipedia 作为数据源，首先通过远程监督方式生成，然后通过手工去除噪声数据。最终 FewRel 1.0 数据集包含 100 个关系，每个关系有 700 个实例。每个句子的平均 token 数量为 24.99，共有 124577 个唯一标记。100 个关系类被分为三部分，其中 64 个关系类用于训练，16 个用于验证和 20 个用于测试。FewRel 2.0 沿用 1.0 的训练集，再此基础上增加生物医学文献数据库 PubMed 作为测试集，共包含 25 个关系类别，每个关系类别有 100 个实例，同时采用 SemEval-2010 任务 8[21]作为验证集。表 1 描述了 FewRel 数据集中的数据格式，其中包括关系 ID、样本语句中包含的单词(tokens)、头尾实体及其位置的标注。

实验部分的样本设置以 C-way-K-shot 任务为标准，具体样本案例如表 2 所示，鉴于篇幅有限仅描述 2-way-1-shot 样本实例，其中蓝色字体代表头实体，红色代表尾实体，训练与测试阶段的样本分别来自 FewRel 1.0 训练集和 FewRel 2.0 测试集，本文主要研究 4 种小样本学习配置，即 5-way-1-shot、5-way-5-shot、10-way-1-shot、10-way-5-shot。实验给出的所有结果均为 10 次训练重复的平均值和标准差，并使用 20000 个独立样本进行测试。

表 1 FewRel 数据集中的数据格式

Tab. 1 Data format in fewrel dataset

key	value
实体关系 ID	P2094
句中包含的单词	["Sasakul", "turned", "pro", "in", "1991", "and", "captured", "the", "WBC", "and", "lineal", "flyweight", "titles", "with", "a", "win", "over", "Yuri", "Arbachakov", "in", "1997", "."]
头实体及位置	["yuri Arbachakov", "Q542462", [文献[17, 18]]]
尾实体及位置	["flyweight", "Q508484", [文献[11]]]

表 2 FewRel 数据集样本设置案例

Tab. 2 Fewrel dataset sample setting case

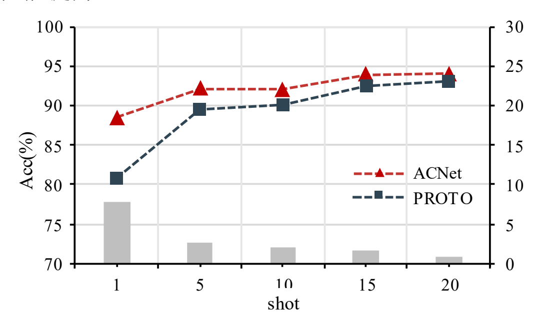
阶段	数据集	设置	案例
training phase	support set	(A)capital of	*Washington* is the capital of *the U.S.A.*
		(B)member of	*Leibniz* was a member of *the Prussian Academy of Sciences.*
	query set	(A)or(B)	*Newton* served as the president of *the Royal Society.*
test phase	support set	(A)inheritance type of	*Hypohidrotic ectodermal dysplasia* is the most common type and is usually transmitted as an *x-linked* recessive trait.
		(B)occurs in	*Congenital* fxi deficiency (*hemophilia c*) is a rare bleeding disorder that has been documented mostly in ashkenazi jews ."
	query set	(A)or(B)	*Acro-dermato-ungual-lacrimal-tooth* syndrome is inherited as an *autosomal dominant* condition .

4.2　实验验证

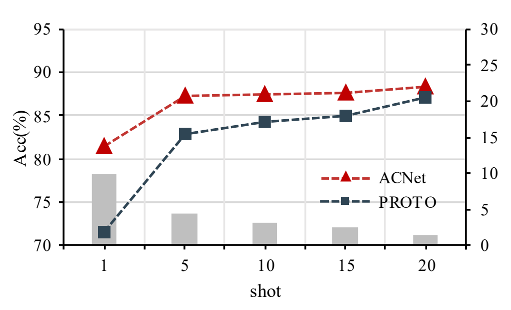
为了验证提出的 ACNet 模型在小样本 RE 任务中的有效性，并评估不同模块对网络的贡献，进行以下三组对比实验。

1) 以原型网络(PROTO)为基线模型，与所提出的模型进行对比，为了保证实验的准确性，使用 BERT 替换 PROTO 中的编码模块。实验中的参数均保持一致，采用 FewRel 1.0 的测试集，分别在 5-way-k-shot 和 10-way-k-shot 任务上进行实验，实验结果如图 2 所示。

图 2(a)和(b)展示了 5-way 和 10-way 任务下，随着 shot 数量增长两个模型准确率的变化曲线，其中灰色柱状图部分代表 ACNet 对比基线模型的提升效果。由图可知，在 K=1 的极端小样本环境下，模型的提升效果最为明显，5-way 任务中相比 PROTO 性能提升 7.83%，10-way 任务上提升 6.27%。PROTO 通常需要一定数量的样本来确定同类样本的质心，并以此作为该类的原型向量，样本数量较少的情况下类原型难以准确反映类别特征，而在 ACNet 中记忆矩阵包含 C_{base} 个类原型，模型在训练期间通过学习类原型的共性，实现样本由部分到整体的归纳。同时，记忆矩阵会在训练过程中对类原型向量进行更新，因此随着样本量的增加，依旧存在一定数量的性能提升。



(a)5-way-k-shot 任务 ACNet 和 PROTO 准确率变化曲线



(b)10-way-k-shot 任务 ACNet 和 PROTO 准确率变化曲线

图 2　ACNet 和 PROTO 在 5--ay 和 10-way 任务上的准确率(%)折线图

Fig. 2　Accuracy (%)line chart of acnet and PROTO on 5 -way and 10-way tasks

2) 以 CapsNet 为基线模型，将路由次数 r 设置为 3 和 5，在 FewRel 1.0 训练集中随机抽取 5000 个样本进行训练。图 3 记录了 CapsNet 在两种路由次数设置下的损失曲线，可以观察到两种配置下的 CapsNet 均达到了系统级收敛。图 4 反映了单个样本在不同路由次数下的 NAS 曲线，当 r=3 或 5 时 NAS 曲线仍成下降趋势，这证明预先设置的路由迭代次数虽然能够使模型满足系统级收敛，但却难以满足模型在实例级的收敛需求，这增加了路由过程的不确定性。

与 CapsNet 不同，ACNet 通过引入的 NAS 函数来判断样本所需的路由迭代次数，图 5 记录了在 10 way-1 shot 任务设置下，达到 NAS 收敛标准时所需的路由迭代次数，图中的灰色和黑色水平线分别代表 3 次和 5 次的迭代次数设置，可以观察到不同类样本所需的迭代次数并不相同，最高需要 9 次路由迭代，最低则仅需要 2 次。数据多样性造成了收敛标准的差异，ACNet 的优势在于针对不同样本的 NAS 分数，自适应调整路由迭代次数，在保证模型整体收敛的同时，满足在实例样本上的拟合标准，从而降低路由过程的不可控风险。

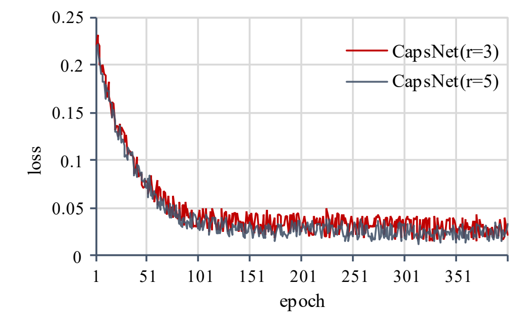


图 3　两种不同路由迭代下 CapsNet 的 loss 曲线

Fig. 3　Loss curve of capsnet under two different routing iterations

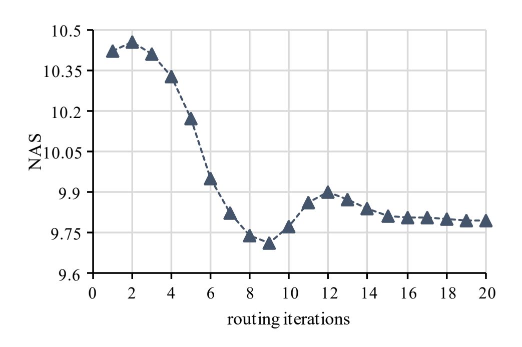


图 4　实例样本在不同路由次数下的 NAS 曲线

Fig. 4　NAS curves of instance samples under different routing iterations

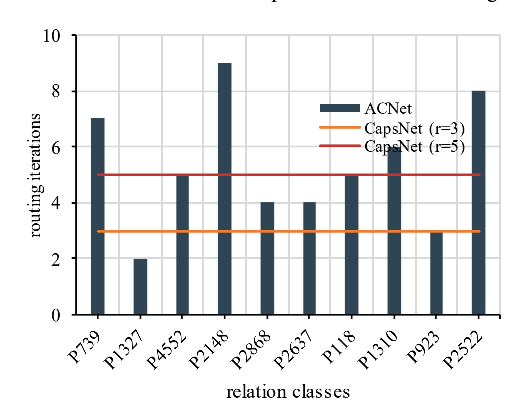


图 5　ACNet 在不同关系类下路由次数

Fig. 5　Routing iterations of acnet under different relationship classes

3) 以原型网络为基线模型，对比 PROTO 增加 QEM 模块前后和 ACNet 模型删除 QEM 模块前后的性能变化。PROTO 在添加 QEM 模块后，将支持集样本和查询样本输入 QEM 模块中，以此获得每个类别的类特征。同时，ACNet 在删除 QEM 后采用取平均的方式获得类特征，其余部分均保持不变。实验使用 5-way-1-shot 任务配置，并采用 FewRel 1.0 测试集进行验证。

如表 3 所示，PROTO+QEM 对比 PROTO 性能提升 4.21%，PROTO 通过类内所有样本向量的加和平均获得类特征，当类内样本较少时，其向量分布存在偏差，无法准确代表样本类别，这进一步影响了查询样本的分类，造成准确率下降。ACNet-QEM 对比 ACNet 精度下降 2.07%，这证明 QEM 具有比 PROTO 更强的样本归纳能力，尽管 PROTO+QEM 采用 QEM 模块获取类特征，但 ACNet-QEM 的性能依旧强于 PROTO+QEM，上升幅度在 13%左右，这证明类特征的编码方式并不是左右性能的唯一指标，本文认为 AIM 模块对结果的影响更加明显，AIM 旨在使用记忆矩阵对支持集向量进行调整，记忆矩阵的引入使得 ACNet 在不同元任务间切换时有效保留学习经验，模块自适应的特性也更加契合小样本学习，实验证明了 AIM 和 QEM 的有效性。

表 3 5-way-1-shot 任务下不同模型的准确率

Tab. 3 Accuracy rate of different models under 5-way-1-shot task /%

序号	模型	5 way-1 shot
1	PROTO	69.20
2	PROTO+QEM	73.41
3	ACNet-QEM	86.44
4	(AIM+QEM)ACNet	**88.51**

4.3 模型比较

4.2 节验证了 ACNet 中不同模块的作用，为了进一步检查 ACNet 的整体模型表现，本文对比了近几年几种常用的小样本基准模型，其中包括：

Meta-Net，Munkhdalai T 等人[8]于 2017 年提出元网络模型，Meta-Net 包含基学习器和带有记忆模块的元学习器两部分，通过两部分的交互驱动模型理解非目标任务空间，减少模型对样本数量的需求。

GNN，Garcia V 等人[22]于 2017 年提出的小样本图卷积网络模型，GNN 将每个支持实例或查询实例视为图中的一个节点，并将其标签信息嵌入到节点表示中，依靠图卷积将基类别的分类器信息传递给新类别的分类器中，实现小样本的标签传播。

SNAIL，Mishra 等人[23]于 2018 年提出的一种元学习框架，SNAIL 将时序卷积网络和注意力机制相结合，利用时序卷积从已有经验中挑选特定信息特征，并通过注意力机制完成信息聚合，达到快速学习小样本任务的目的。

PROTO，Snell 等人[10]于 2017 年提出原型网络，PROTO 通过平均支持样本获得样本的类别中心，并将新样本与类别中心进行距离度量，以此实现少量样本下的分类任务。

BERT-PAIR，Tianyu Gao 等人[20]于 2019 年提出一种序列匹配模型，BERT-PAIR 将每个查询实例与所有支持实例进行配对，并将每个实例对连接为序列输入到 BERT 模型中，获得表示同类实例的概率。

DBIN，Ruiying Geng 等人[24]于 2020 年提出动态内存引导网络，DBIN 将内存模块和动态路由算法相结合，模型通过调整和聚合两个步骤在少量支持样本中获得类别表示，最后与查询样本比较完成关系分类。

实验结果如表 4 所示，其中 Meta-Net、GNN、SNAIL、PROTO(CNN)等网络模型使用 CNN 进行编码，输入关系样本语句，并将每个单词表示转换为单词嵌入和位置嵌入的整合，整个关系实例表示作为输入向量。在 CNN 中输入向量经过卷积层、最大池化层和非线性激活层得到最终的语句嵌入。除了使用 CNN 编码结构外，PROTO(BERT)、BERT-PAIR、DMIN 以及本文提出的 ACNet 模型均采用 BERT 作为编码器。从表一中可以看出 PROTO(BERT)对比 PROTO(CNN)存在显著的性能提升，在 FewRel 1.0 数据集的 4 个任务上平均提升幅度为 4.72%，在 FewRel 2.0 数据集上，平均的性能提升幅度为 3.09%，这证明 BERT 编码结构所生成的语句特征更加丰富，能够有效应对小样本任务。

同时，对比 FewRel 1.0 数据集，所有的模型在 FewRel 2.0 数据集上都存在性能大幅降低的现象，FewRel 2.0 的测试集来自生物医学领域，这说明小样本模型难以快速适应跨领域样本，模型的经验迁移能力还存在较大的提升空间。

表 4 不同模型在四个小样本任务设置上的对比

Tab. 4 Comparison of different models in four few-shot task settings

模型	5-way-1-shot		5-way-5-shot		10-way-1-shot		10-way-5-shot	
	On 1.0	On 2.0	On 1.0	On 2.0	On 1.0	On 2.0	On 1.0	On 2.0
Meta-Net (CNN)	64.46	-	80.57	-	53.96	-	69.23	-
GNN(CNN)	66.23	27.94	81.28	29.33	46.27	16.44	64.02	18.26
SNAIL(CNN)	66.79	26.22	79.04	30.28	45.73	16.21	68.33	19.36
PROTO(CNN)	74.52	35.09	88.40	49.37	62.38	22.98	80.45	35.22
PROTO(BERT)	80.68	40.12	89.60	51.50	71.48	26.45	82.89	36.93
DMIN	85.14	49.62	92.37	53.76	76.56	40.78	86.75	47.49
BERT-PAIR	88.32	56.25	93.22	67.44	80.63	43.64	87.02	53.17
ACNet	**88.51**	**58.44**	**93.49**	66.53	**81.22**	**45.74**	**87.32**	52.24

在 FewRel 1.0 数据集上，本文提出的 ACNet 模型优于目前最优的 BERT-PAIR 模型，在 FewRel 2.0 的两种 one-shot 任务上，对比 BERT-PAIR 也分别取得了 2.19%和 2.1%的提升，而在其他的两种 5-shot 任务上，ACNet 也达到了与 BERT-PAIR 相似的性能。BERT-PAIR 是一种基于匹配的小样本方法，通过计算查询集与支持集的匹配程度完成分类，但鉴于样本的特征多样性，同类样本的分布差异较大时会导致模型性能下降，而 ACNet 对支持集样本采用先调整后聚合的方式，SIA 针对每个样本进行路由评估，调整后的样本能够有效减少类内分布差异，自适应的特性使样本在实例级别上获得收敛，因此面对跨领域的样本，ACNet 仍具有一定的优势。

综合本节中的实验结果，本文提出的 ACNet 模型在 FewRel 数据集的多数任务上都优于目前模型，实验验证了本文方法的有效性，同时也说明类特征建模的方法在小样本任务中的优势。

5 结束语

本文提出了一种自适应胶囊网络(ACNet)，并将其应用在小样本关系分类中。ACNet 利用记忆矩阵帮助模型归纳类表示，并通过自适应归纳算法完成对支持集向量自适应调整，使模型能够发现新的未知类。在 FewRel 数据集上，与当前五个代表性模型相比，在 FewRel 1.0 的 4 中小样本任务上取得了最好的结果，而在 FewRel 2.0 的两种 10-way 任务上具有与目前最优模型 BEER-PAIR 的相似性能。目前 ACNet 还存在模型结构复杂，训练时间长等问题，在未来的工作中，本文将进一步研究模型规模对小样本关系抽取任务的影响，探索精简模型结构的可能。

参考文献：

[1] 杨穗珠，刘艳霞，张凯文，等. 远程监督关系抽取综述 [J]. 计算机学报, 2021, 44 (08): 1636-1660.

[2] 刘洋. 神经机器翻译前沿进展 [J]. 计算机研究与发展, 2017, 54 (06): 1144-1149.

[3] 付雷杰，曹岩，白瑀，等. 国内垂直领域知识图谱发展现状与展望 [J]. 计算机应用研究, 2021.

[4] Han Xu, Gao Tianyu, Lin Yankai, *et al.* "More Data, More Relations, More Context and More Openness: A Review and Outlook for Relation Extraction. " [C]// Proceedings of the 1st Conference of the Asia-Pacific

Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, 2020: 745–758.

[5] Salamon J, Bello J. Deep convolutional neural networks and data augmentation for environmental sound classification [J]. IEEE Signal processing letters, 2017, 24 (3): 279-283.

[6] Huisman M, VanRijn J, Plaat A. A survey of deep meta-learning [J]. Artificial Intelligence Review, 2021, 54 (6): 4483-4541.

[7] Qu M, Gao T, Xhonneux L P, *et al.* Few-shot relation extraction via bayesian meta-learning on relation graphs [C]// International Conference on Machine Learning. PMLR, 2020: 7867-7876.

[8] Munkhdalai T, Yu Hang. Meta networks. [C]// In ICML'17 Proceedings of the 34th International Conference on Machine Learning-Volume 70, 2017: 2554–2563.

[9] Qi Hang, Brown M, Lowe D. Low-Shot Learning with Imprinted Weights. [C]// In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 5822–5830.

[10] Snell J, Swersky K, Zemel R. Prototypical Networks for Few-Shot Learning [J]. In Advances in Neural Information Processing Systems, 2017, 30: 4077–87.

[11] Sung F, Yang Yongxin, Zhang Li, *et al.* Learning to compare: Relation network for few-shot learning [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1199-1208.

[12] Sabour S, Frosst N, Hinton G. Dynamic Routing Between Capsules [J]. In Advances in Neural Information Processing Systems, 2017, 30: 3856–66.

[13] Devlin J, Chang Mingwei, Lee K, *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv: 1810. 04805, 2018.

[14] Vaswani A, Shazeer N, Parmar N *et al.* Attention Is All You Need. [C]// In Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, 30: 5998–6008.

[15] Yang Zhengxin, Zhang Jinchao, Meng Fandong, *et al.* Enhancing Context Modeling with a Query-Guided Capsule Network for Document-Level Translation. [C]// In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) , 2019: 1527–37.

[16] Jones M, Keynes M. On kernel density derivative estimation [J]. Communications in Statistics-Theory and Methods, 1994, 23 (8): 2133-2139.

[17] Comaniciu D, Meer P. Mean Shift: A Robust Approach toward Feature Space Analysis. [C]// IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002, 24 (5): 603–19.

[18] Vinyals O, Blundell C, Lillicrap T, *et al.* Matching networks for one shot learning [J]. Advances in neural information processing systems, 2016, 29: 3630-3638.

[19] Han Xu, Zhu Hao, Yu Pengfei, *et al.* FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation [J]. arXiv preprint arXiv: 1810. 10147, 2018.

[20] Gao Tianyu, Han Xu, Zhu Hao, *et al.* FewRel 2. 0: Towards more challenging few-shot relation classification [J]. arXiv preprint arXiv: 1910. 07124, 2019.

[21] Hendrickx I, Kim S, Kozareva Z, *et al.* SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals. [C]// In Proceedings of the 5th International Workshop on Semantic Evaluation, 2010: 33–38.

[22] Garcia V, Bruna J. Few-shot learning with graph neural networks [J]. arXiv preprint arXiv: 1711. 04043, 2017.

[23] Mishra N, Rohaninejad M, Chen Xi, *et al.* A simple neural attentive meta-learner [J]. arXiv preprint arXiv: 1707. 03141, 2017.

[24] Geng Ruiying, Li Binhua, Li Yongbin, *et al.* Dynamic memory induction networks for few-shot text classification [J]. arXiv preprint arXiv: 2005. 05727 (2020) .